Extending Digital Repository Architectures to Support Disk Image Preservation and Access

Kam Woods
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, CB #3360
1-(919)-966-3598
kamwoods@email.unc.edu

Christopher A. Lee
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, CB #3360
1-(919)-962-7204
callee@ils.unc.edu

Simson Garfinkel
Graduate School of Operational and Information Sciences,
Naval Postgraduate School
Monterey, CA
1-(831)-656-3389
slgarfin@nps.edu

ABSTRACT
Disk images (bitstreams extracted from physical media) can play an essential role in the acquisition and management of digital collections by serving as containers that support data integrity and chain of custody, while ensuring continued access to the underlying bits without depending on physical carriers. Widely used today by practitioners of digital forensics, disk images can serve as baselines for comparison for digital preservation activities, as they provide fail-safe mechanisms when curatorial actions make unexpected changes to data; enable access to potentially valuable data that resides below the file system level; and provide options for future analysis. We discuss established digital forensics techniques for acquiring, preserving and annotating disk images, provide examples from both research and educational collections, and describe specific forensic tools and techniques, including an object-oriented data packaging framework called the Advanced Forensic Format (AFF) and the Digital Forensics XML (DFXML) metadata representation.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries—*collection, dissemination, systems issues.*

General Terms
Archiving, Digital Forensics, Disk Images.

Keywords
Forensic Datasets; Digital Forensics XML (DFXML); Advanced Forensic Format (AFF); Long-Term Digital Preservation.

1. INTRODUCTION
Much of the literature on digital archives emphasizes the "virtual" (i.e. intangible) nature of electronic resources. Computer systems have "an illusion of immateriality by detecting error and correcting it" [21]. However, digital objects are created and perpetuated through physical phenomena (e.g. charged magnetic particles, pulses of light, pits in disks). This materiality brings challenges: data must be read from specific artifacts. Those artifacts can become damaged or obsolete, resulting in partial or complete loss of access to data.

The materiality of digital resources also brings unprecedented opportunities for description, interpretation and use [22]. For example, instead of merely studying a final letter, a researcher could analyze data associated with the authoring process and characteristics of the creator's working environment, in order to test assumptions about the letter's provenance and better understand its context of creation and use.

A fundamental mechanism for enabling the curation of underlying data is the acquisition and management of *disk images*—sector-by-sector copies of the data from physical storage media, including modern hard drives, optical disks, USB storage devices, and even portable devices such as iPods, digital cameras, and mobile phones. In this paper, we discuss a variety of issues and opportunities for extending digital repository architectures in order to treat disk images as objects within collections.

2. DISK IMAGES AS DIGITAL OBJECTS
Thibodeau [31] describes every digital object as simultaneously being:

1. A *physical object*, or an "inscription of signs on some physical medium."
2. A *logical object*, or a digital artifact that is "recognized and processed by software."
3. A *conceptual object*, or one that is "recognized and understood by a person, or in some cases recognized and processed by a computer application capable of executing business transactions."

Attempting to reproduce data at different levels of representation requires specialized methods and tools. These may range from simply copying the user-accessible files within the file system, to recovery of overwritten or deleted data, to the assemblage of metadata and use of statistical methods to determine access profiles and identify users associated with particular data items [15].

Conventionally access to data on a storage device involves mounting a volume. The operating system, and in particular the file system, mediates access to the underlying data. The file system allocates names to groups of sectors (e.g. filenames), and

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUN 2011		**00-00-2011 to 00-00-2011**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Extending Digital Repository Architectures to Support Disk Image Preservation and Access	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Naval Postgraduate School,Graduate School of Operational and Information Sciences,Monterey,CA,93943**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Disk images (bitstreams extracted from physical media) can play an essential role in the acquisition and management of digital collections by serving as containers that support data integrity and chain of custody, while ensuring continued access to the underlying bits without depending on physical carriers. Widely used today by practitioners of digital forensics, disk images can serve as baselines for comparison for digital preservation activities, as they provide fail-safe mechanisms when curatorial actions make unexpected changes to data; enable access to potentially valuable data that resides below the file system level and provide options for future analysis. We discuss established digital forensics techniques for acquiring, preserving and annotating disk images, provide examples from both research and educational collections, and describe specific forensic tools and techniques, including an object-oriented data packaging framework called the Advanced Forensic Format (AFF) and the Digital Forensics XML (DFXML) metadata representation.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**10**	

allows files to be grouped together in "folders" or "directories." Simply put, the file system decides "where and how it stores information" [10]. Disks also contain information that is not immediately apparent to the casual user—such as hidden or deleted files, alternate partitions, configuration settings, documentation of operations performed on the computer, and data that support cross-platform compatibility.

2.1 Acquiring Disk Images

Forensic investigations are typically conducted on *images* of data captured from primary sources rather than on the original media. Using an image assures that the original media will not be inadvertently modified or otherwise compromised—a concern shared by both forensic investigators and digital curation professionals.

A disk image is a sector-by-sector copy of the data that was stored on a physical medium. As such, the disk image is a "snapshot" of the medium's content, including all allocated files, file names, and other metadata information associated with the disk volume. Once a disk image has been generated, it is then stored as a single file or set of files. The disk image files serve as the most general of containers, because they can contain anything that has been stored on a computer.

Disk images allow researchers to retain and investigate aspects of the systems that could be inadvertently altered during normal operation of a typical operating system. For example, the mere act of turning on a computer and booting the operating system, or moving an external storage device from one computer to another, can result in data being changed and possibly destroyed. These changes may include modifications to operational metadata and other aspects of the original data objects such as byte order, character encoding of specific objects, file system information, MAC (modified, accessed, created/changed) values, access permissions, and file sizes. The risk of losing such information can be exacerbated by attempting to access original removable or fixed media originally produced on or by another operating system, such as when using a Windows machine to access HFS-formatted Macintosh drives.

Because of these potential pitfalls, disk images are created using special-purpose tools that access the physical device with low-level input-output operations without using the host computer's file system as an intermediary. Hardware *write blockers* are used to ensure that source devices are not inadvertently altered or contaminated during capture.

Traditionally, disk images were created with the UNIX tool "dd." Today it is more common for forensic investigators to create images using tools that additionally record metadata (e.g. the name of the investigator, time that the image was created, and notes) and integrity information (e.g. checksums or hashes). Disk images can be created by the free commercial application FTK Imager or by the Disk Utility that is included with the Macintosh operating system.

2.2 Residual Information and Viruses

Disk images may contain *residual data* from previous use of the computer system. For example, a disk image may include individual files that have been deleted, are no longer visible using the file system, but which can still be recovered. Also present may be fragments of files that have been partially overwritten, and memory artifacts from hibernation files and virtual memory activity. This residual data can be valuable in reconstructing or making inferences about an individual's prior activities, state of mind, intentionality, as well as technical information such as how a computer was used, devices previously connected to it, and other computers on a network with which it interacted.

There are many reasons why it can be desirable for collecting institutions to identify "hidden" information associated with materials under their care . For example, a producer or donor may inform the institution that valuable documents have accidentally been deleted or "lost." The producer could also be an organization with a policy of encrypting files that are to be released publicly at a later date, but has lost the encryption key or passphrase. The creators of the materials could be a group of individuals who wish to include specific information on authorship or attribution, but do not know who was logged on to a particular machine at a particular time. System information–including the Windows registry, log files, hibernation and backup data—can often be used to recover relevant data to address such issues. Various forms of system information can also help to identify and resolve system dependencies, by revealing what applications were installed and run on the computer, and the software was configured and used. Timestamp information associated with system files can help to reconstruct the chronology of activities on a machine in ways not possible using solely the timestamps of data files. Information about previously mounted devices, network connections and internet activities can alert collecting institutions to additional sources of valuable information (e.g., a scientist's laptop that reveals she stored all of her researcher data on an external hard drive or a photographer's computer that reveals he stored his work using an hosted service on the Web).

A disk image can serve as a kind of sandbox or staging area. As we discuss later in this paper, there are a variety of ways to expose users to content from the image without requiring them to mount the drive or run the original file system. This can help to minimize the risk of infecting users' computers with legacy computer viruses.

2.3 Retention and Packaging of Forensic Data

Once a disk image is created, it can be stored as a single object in a repository. Because modern disk images are compressed, they are typically one-half to one-sixth the size of the source media. For example, imaging a 250GB laptop hard drive would typically result in a digital artifact ranging between 50GB and 150GB in size. This disk image could be a single file, or it could be split into 25 to 75 sequentially numbered files of roughly 2GB each.

The EnCase Evidence File Format (LEF or E01) and Advanced Forensic Format (AFF) both serve as packages for disk images. In addition to the raw disk image data, they include metadata to ensure both proof of file integrity and to document chain of custody. By validating the hash value of a disk image, one can establish that the bitstream obtained from the original medium has not been altered. This can assure that digital object characteristics, properties and associated metadata will not be lost, as long as a repository can continue to interpret the disk image packaging formats. By contrast, file-level treatment of digital collections can introduce various forms of irreversible data loss. Digital repositories may be better served in many cases by assuring the

integrity of entire bitstreams, rather than attempting to guarantee at the time of ingest that they will be able to render and reproduce all essential file-level characteristics [27].

One can also record and verify hash values for files of interest, and file system metadata can frequently be used to determine if a particular user has interacted with a given file on a given date or over a given period. As a whole, these types of data allow one to describe the "ecology" of a particular drive, extracting and analyzing the types of contextual information relevant to many archival processes [23]. Block-level checksums provided by forensic image formats further support recovery or repair of damaged disk images.

Processing and interpretation of more complex drive compositions can likewise be performed using modern forensic file formats. One implementation of AFF provides a mechanism to store multi-volume RAID setups as volume streams within a single AFF object; the process by which the RAID mapping can be reconstructed and written back to the AFF object is described in [7].

The retention and preservation of disk images can be beneficial from a risk management perspective. The most common disk image packaging formats are well documented, and extraction of the raw bitstream can be performed by a variety of existing tools. Likewise, open forensic disk formats such as AFF store and produce metadata that conforms to a standard that can be easily parsed and mapped or appended to existing preservation metadata schemas. This is discussed further in section 4.1.3.

2.4 Working with Disk Images

The most straightforward way for a researcher to access information stored in a disk image is to mount it as a virtual device. This allows the researcher to "browse" the data contained in the image as if the original physical device were connected to the researcher's computer.

Accessing disk images via a host mount imposes a number of technical limitations. In particular, the researcher does not have access to deleted files or partially overwritten information, and may incur security risks on the host due to virus infections present on the imaged system.

Individual files within the disk image can be listed or extracted using a variety of free and commercial digital forensics tools. Using these tools, curators of collections and their users can also perform keyword searches through documents (including compound documents such as Adobe Acrobat and Microsoft Word files), identify encrypted files, recover deleted files, and create timelines of users' activity (as evidenced by file modification timestamps).

While the techniques, terms, and immediate motivations for digital forensics investigations differ significantly from those in typical digital archival activities, many of the fundamental goals align: reducing risk to the data source in acquisition and analysis, mindfulness of sensitive and confidential data, and maintaining a record of actions performed during handling activities.

2.5 Disk Images and Repositories

Disk images can play an essential role in the acquisition and management of digital collections [8][18][34]. Preserved disk images can be used at a later time to provide proof of file integrity and chain of custody. Disk images can ensure continued access to information in collections without depending on physical carriers, which may be fragile or become obsolete; can serve as baselines for comparison when evaluating digital preservation actions; and can provide fail-safe mechanisms (backups) when curatorial actions make unexpected changes to data. Disk images can be shared with other institutions. Finally, disk images provide access to potentially valuable data that resides below the user-accessible portions of the file system, including metadata, recoverable sectors and configuration information.

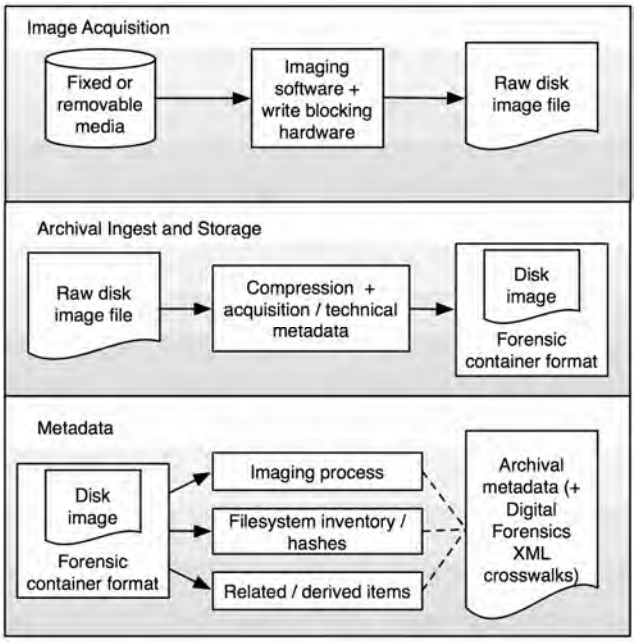

Figure 1: Storage media acquisition and handling profile for digital repositories.

An overview of a disk imaging, storage, and description process is illustrated in Figure 1. Note that many forensic formats support export of metadata related to both the file system and acquisition process.

2.6 Preservation and Access

There are likely to be changes in preservation strategy or access conditions over time. Consider a case in which the default ingest process is to create a normalized Archival Information Package (AIP) from a given type of Submission Information Package (SIP)—*e.g.,* convert all Word documents to PDF/A. Even in controlled conditions, this process will involve some loss of information (either in the file formatting itself or through original metadata which is no longer directly linked to the file). Future techniques or access scenarios might require access to the original Word files—and possibly also information embedded in the file system—in order to recover data not present in the final rendition.

Depending on understanding of arrangement with the Producer, hidden data can also serve as a valuable resource for the curation of a collection. Repository professionals could make use of traces of data that indicate what application created files, and login or password information necessary to access various sources, including online data stores that were accessed by the computer's user [12].

When caring for digital collections, it is important to be cautious about making *irreversible transformations*. Copying files off of the original medium and then discarding that bitstream is just such an irreversible change, because there is no way to then recover the original bits from the extracted files. Inadvertent or intentional alteration of files and file attribution is also a common side-effect of some archival procedures.

Disk imaging can thus dramatically reduce the possibility that such changes are irreversible, by assuring that bit-perfect representations of the original media will be preserved in modern storage systems. It is important to clarify that disk imaging is not simply a means to *postpone* future transformations of the data, nor does it require digital archives to adopt emulation of the original computing environment as their digital preservation strategy. Rather, it supports recognition of the fact that objects with complex structural, semantic, and relational properties should (when possible) be preserved in a manner that supports a high degree of flexibility in future access and analysis.

3. BUILDING DISK-IMAGE COLLECTIONS

Several recent initiatives have focused on building collections of disk images. These are in response to a number of critical factors: (1) a recognition that large legacy collections of removable media held at various institutions are at risk not only due to physical degradation or failure, but also due to limitations of current processing and analysis capabilities, (2) an influx of physical media and media images from external sources, and (3) a recognition of historical losses that have been incurred in previous attempts to handle digital materials.

3.1 Acquisition by Collecting Institutions

Recovery of data from physical media has been an occasional topic of discussion in the professional library and archives literature [28][35]. For several years, the Cornell University Library ran a File Format & Media Migration Service [9], which focused on recovery of data from obsolete or at-risk media. A project at Indiana University built a collection of disk images from media distributed by the United States Government Printing Office and developed mechanisms for managing and providing access to the images [33][34].

Several authors have recently investigated the use of forensic tools and techniques for acquiring digital collections in libraries and archives [12][18]. The Prometheus [8] and PERPOS [31][32] projects have developed software for data extraction, focusing on needs of specific collecting contexts. The e-Depot environment has been designed to accommodate workflows that include acquisition of disk images [26]. Born Digital Collections: An Inter-Institutional Model for Stewardship (AIMS) and FutureArch are exploring workflows that include forensic copies from digital media. Projects called “Digital Lives” and “Computer Forensics and Born-Digital Content in Cultural Heritage Collections” have provided significant contributions to this discussion [19][21].

Several collecting institutions have been actively testing and applying digital forensics methods and building forensic lab environments for their acquisitions, including the Bodleian, British Library, City of Vancouver, Emory University, and Stanford University Libraries. At the University of North Carolina at Chapel Hill (UNC-CH), we are incorporating forensic tools and processes into educational initiatives to provide students, archivists, and curation experts with the knowledge and skills necessary to address the rapidly evolving needs of the profession.

The capture and preservation of disk or memory images also plays an important role in the preservation of computer game collections. The inclusion of these images in repositories requires adaptation and extension of existing digital library conventions for representing objects and associated metadata [25]. The CAMiLEON project tested user interactions with different instantiations of a video game called Chuckie Egg, which ran on the BBC Micro: running natively on the original BBC Micro platform, a disk image used in a Windows environment using an emulator, and a version that was migrated to run directly in Windows [17]. The first two Digital Preservation Challenges, administered in 2008 and 2009 by DigitalPreservationEurope also included memory images as objects to be addressed by challenge contestants.

3.2 Forensic Data Corpora

Currently available forensic corpora provide valuable support for research, education, and training of both forensics investigators and digital curation professionals.

3.2.1 Real Data Corpus

The Real Data Corpus consists of more than 2,000 used hard drives purchased on the secondary market for the purpose of cutting-edge forensics research [13].

This corpus was assembled for the purpose of developing and validating forensic and data recovery tools, training students in forensics and data recovery, and additional research into document analysis and transformation. Because the data is drawn from real entities and devices (including personal and corporate computers, ATMs, and medical and industrial equipment), this corpus provides a unique window into real-world computing practices both at the level of the user and in terms of security, configuration, and long-term management. Unfortunately, because these disks contain information that is private and sensitive, they are inappropriate for use in training and education scenarios.

3.2.2 “M57 Patents:” A Realistic Corpus for Education

Educators require corpora that simulate real-world data but that do not contain information that is private or sensitive. However, it is challenging to create data sets that are complex enough to provide genuine challenges in classroom environments.

In this section we present the “M57 Patents” corpus, a scenario constructed for the purpose of enabling digital forensics education and research. The corpus was created by student researchers on a private network of computers over a period of 18 days. Different students played the role of different individuals as the start-up company experienced a variety of criminal scenarios including the exfiltration of corporate data, theft of equipment, and the storage of illegal digital materials.

The corpus is comprised of daily disk images, network captures, and RAM dumps from workstations. The corpus additionally includes USB disk images and a memory image from a cellular telephone. These materials total more than 600GB of data.

Because the design and implementation of the corpus focuses on realistic evolution of the file systems over time (simulating real-

world "wear" on a system as it ages), this corpus is also well-suited for use in the education of information professionals who are likely to acquire media that contain materials generated over extended periods, as well as the creation of tools for such professionals to use. Commercial and open source tools to perform automated tasks such as bulk data extraction, triage to identify relevant information based on both simple and contextual filters, and generate custom reports are widely used in the forensics community.

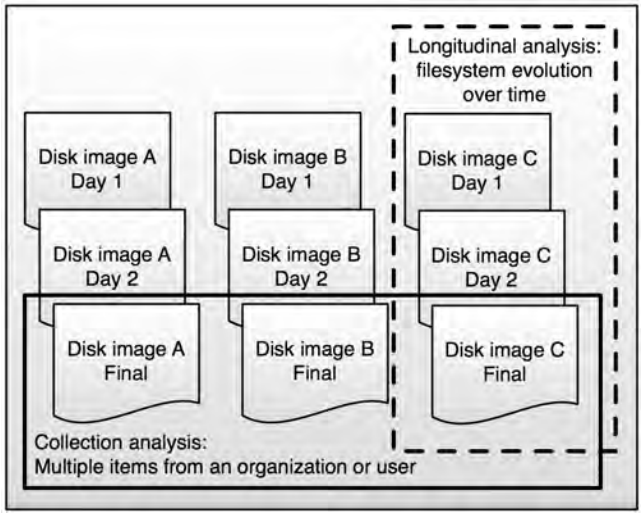

Figure 2: Modes of disk image collection analysis.

Realistic corpora such as the M57 Patents corpus can also be used to describe and assist in the improvement of mechanisms for digital curation, along with distribution of large disk image corpora. As part of this work, we have worked with iBiblio at UNC-CH to provide BitTorrent access to this collection;[1] torrents allow us to create "views" into the corpus that draw on particular bitstreams from the underlying storage—for example, all of the RAM images, or only the disk images created on the final day of the scenario. Two such views are illustrated in Figure 2.

There are numerous advantages to using BitTorrent to disseminate the images. Practicing educators and researchers who wish to use the data benefit from a distribution system supported by their peers, and can elect to use only those portions of the (freely available) corpus that are relevant to their activities. The data were generated in ways that minimize concerns about copyright protections, in order to support both *data sharing* with respect to experimental work, and *tool validation*—that is, researchers and educators have a "ground truth" for what is contained in the corpus and can use this to explicate the operation (or failure) of various software packages used for analysis.

In addition to ongoing research identifying forensics practices that are relevant to digital curation, we have incorporated data from the M57 Patents corpus into two class at UNC-CH to expose students to methods for handling large-scale disk image collections and working with commercial and open-source forensics and bitstream analysis tools.

[1] http://digitalcorpora.org/corpora/scenarios

4. ISSUES ASSOCIATED WITH CURATION OF DISK IMAGE COLLECTIONS

Curation of forensic data can be complex, as forensic investigations often incorporate data captured from multiple sources. This is similar to many acquisitions of collecting institutions, which aggregate a diversity of materials with a common provenance (individual, function, process or organization). Curation of forensic data in collections will require reliable and accurate mechanisms for dealing with provenance, metadata consistency (of files and between files and file systems), bit accuracy, and performance requirements for access and analysis.

4.1 Technical Challenges

Raw bitstreams such as those generated during forensic disk imaging are generally produced using tools that output technical, administrative, and rights metadata. Metadata associated with a raw stream will typically include low-level hardware information such as drive geometry, system information from the capture hardware, and cryptographic hashes to verify integrity. High-level metadata such as that found in a typical Submission Information Package (SIP) must be generated after the fact. Contextual links between any captured drives and supporting materials can be produced via a secondary process, either specified during capture or prior to ingest. Finally, information on image provenance and integrity should be recorded either via a packaging mechanism (such as the "case files" used in forensic investigations) or within formats specifically designed to support the addition of flexible metadata such as the Advanced Forensic Format (described in a later section) [8].

4.1.1 Digital Curation Policies

Born-digital materials in archival practice often undergo various stages of transformation during the sequence of acquisition, ingest, archival storage, and access. The nature of these transformations is typically specified prior to any technical process applied to the objects in question. Digital curation policies for such materials may address media and format migration, rights management, metadata transfer and creation, administrative processes, and access filters, among other concerns.

A primary question is whether to retain the entirety of the raw image over time, or instead to keep the complete image only temporarily as part of an ingest staging area. Such decisions can be affected by many considerations beyond basic storage and processing requirements, although continual drops in the cost of raw storage increasingly enable collecting institutions to avoid procedures that result in permanent information loss.

Long-term planning for retention can depend on cost factors affected by the nature of the media–e.g., whether the ingest process is dealing with collections of low-density floppy disks, multiple-Terabyte hard disks removed from modern computer systems, or a highly heterogeneous collection.

The cultural, historical, scientific, and economic reasons for acquiring and retaining born-digital materials—along with the intended or projected use cases—have a significant effect on both archival planning decisions and development of the technical infrastructure necessary to support retention and access over time.

Important factors include the institution's commitment to donors and other stakeholders, and an understanding of any ongoing support that will be required to maintain a given collection.

4.1.2 Improving Preservation Profiles

Digital forensics tools and practices provide numerous options for extracting and preserving digital information from disk images that can be incorporated into future archival practice. *Evidence files*, or container formats and methods used to wrap raw disk images with acquisition, chain-of-custody, and other administrative and technical metadata, are commonplace in forensic practice. They support authenticity, security, and flexible data export (both at the level of individual file objects and for classes of data depending on the tools in use), all factors that are relevant to the curation of digital materials.

Forensic tools also provide support for exporting a wide array of data from raw bitstreams. This is generally done to address some identified need on the part of the end-user (a digital forensics investigator or expert witness). Providing filtered or restructured views into raw disk images is also desirable in a preservation and archival access context.

There is some previous work on specifying XML schemas for forensic data, including XIRAF [1], and Digital Forensics XML [16] provides API-level support for fine-grained analysis and summarization of the bitstream.

4.1.3 Metadata

The extensive technical and file system metadata that can be extracted from raw bitstreams (and which is stored in or alongside forensics container formats) provides multiple paths to creation of the types of technical, administrative, and rights metadata familiar to experts in digital curation. Similarly, recently introduced metadata standards and tools can be used to reflect the rich semantic relationships both between objects within the raw bitstream and between disk images (or versions of the same disk image captured at different times or prepared for different purposes).

The PREMIS data model provides a facility to describe multiple *representations* of a digital object, *events* associated with changes or modifications to that object over its archival lifetime, and *agents* associated with these events [3]. The PREMIS documentation and description of intended uses focuses predominantly on files that can be traditionally rendered (for example, TIFF images). While it is possible to apply this model to disk images stored as files that can be accessed and manipulated using dedicated tools (rather than mounted as file systems), events at multiple levels of representation (those that occurred both within the Producer and Archive systems) may not have a natural mapping.

Forensic disk container formats—particularly AFF—already incorporate significant metadata related to the acquisition, composition, and identification, and location of image data. For disk images in particular, one can identify hashes of the raw disk image within the container format (MD5 and SHA1), the time and date of acquisition, and the acquisition tool (among other values).

Finer granularity can be accomplished through the use of Digital Forensics XML and tools such as *fiwalk* [16]. The *fiwalk* tool can export XML files (in addition to CSV and simple lists) that correspond to particular views of the file system—for example, the name, type, path, and hash value of every Microsoft Word document within the raw disk image. Such XML files can be used to dynamically generate views into the raw data, to support comparison between multiple copies of the same disk image, or to create data streams for export and distribution. Digital Forensics XML (built in part around Dublin Core) can be converted to metadata standards such as METS and EAD.

The use of Digital Forensics XML (DFXML) allows multiple tools to work together, sharing data and work products both for the file system and at levels other than individual files and/or disk images. DFXML is an emerging standard being developed around a common set of tags and data representations (including a DTD and schema for validation).[2]

```
Segment                  arg   length   data
=======              ========  =======  ====
. . .
afflib_version             0        7  "3.3.3"
aff_file_type              0        3  AFF
acquisition_commandline    0       36  aimage /dev/sda
                                         /mnt/charlie-002
acquisition_device         0        8  /dev/sda
sectorsize              1024        0
pagesize               16777216    0
devicesectors              2        8  = 9999064 (64-bit value)
acquisition_macaddr        0       18  00:0b:db:4f:6b:10.
acquisition_dmesg          0    27298  [    0.000000]
                                        Initializing cgro
image_gid                  0       16  7256 F895 DE4F E304
                                        233E 21C0 2347 CCC5
acquisition_date           0       20  2009-11-12 19:12:18.
md5                        0       16  0609 2DFE AA4F B183
                                        946F 95D8 AD84 519E
acquisition_seconds     1570        0  = 00:26:10 (hh:mm:ss)
imagesize                  2        8  = 10239860736 (64-bit value)
```

Figure 3: Partial view of the raw image metadata captured by *aimage* Note the globally unique identifier, acquisition date and time, and hash value.

Metadata captured by digital forensics tools, as illustrated in Figures 3 and 4, can thus be used to enhance profiles provided by preservation metadata schemas at varying levels of specificity. By capturing detailed data from the file system *from which the file originated*, it is possible to provide useful information linking object provenance, object validation, and object transformations performed. Critically, the burden of *retaining* this data does not depend solely on the metadata schema which is employed, or the curation practices implemented at a particular time, as long as one maintains the original bitstream and an ongoing ability to process it.

These records may contain data that convey subtleties not typically recorded in archival descriptions–e.g., *change* versus *modification* times and dates, indicating (respectively) the last alteration within a file system (for example, movement from one directory to another) and the last alteration of the actual file contents. Recording these types of data can assist in eliminating assumptions about what may be of interest to future researchers and historians–without incurring significant storage or processing overhead.

[2] Details about the current state of DFXML can be found at http://www.forensicswiki.org/wiki/Category:Digital_Forensics_XML

```
<fileobject>
      <filename>Documents and Settings/All Users/Documents/
                My Pictures/Sample Pictures/Blue hills.jpg
      </filename>
      ...
      <filesize>28521</filesize>
      <alloc>1</alloc>
      <used>1</used>
      <inode>6245</inode>
      ...
      <uid>0</uid>
      <gid>0</gid>
      <mtime>1208174400</mtime>
      <ctime>1257729636</ctime>
      <atime>1257729636</atime>
      <crtime>1257729636</crtime>
      <seq>2</seq>
      <libmagic>JPEG image data, JFIF standard 1.02</libmagic>
      <byte_runs>
       <run file_offset='0' fs_offset='0' img_offset='363200512'
         len='0'/>
      </byte_runs>
      <hashdigest type='MD5'>
          6fb2a38dc107eacb41cf1656e899cf70
      </hashdigest>
      <hashdigest type='SHA1'>
          4eee44b18576e84de7b163142b537d2fe6231845
      </hashdigest>
</fileobject>
```

Figure 4: Partial view of XML output from *fiwalk*. Items in red include the original file system path, MAC values, file format, and hash values.

4.2 Access, Rights, and Administration

Disk images of commercially-produced materials may be subject only to the distributional and licensing agreements by which they were originally bound. However, archives are increasingly tasked with handling disk images obtained from individual users or organizations, with confidentiality and privacy becoming significant concerns. Such disk images may contain privileged information. Disk images may contain both information intended for public access and protected intellectual property - commercial software, information intended for future sale or licensing, among others.

4.2.1 Confidentiality and Security

Disk images obtained from private parties or trusts may be subject to more complex preservation, legal, and access arrangements—for example closure or limitations of data disclosure until death.

Preserving confidentiality of private data is a significant concern for archives acquiring data from raw disk images. In addition to issues of privileged information identified previously, archivists may be concerned with identifying and redacting or limiting access to digital objects encumbered with intellectual property agreements or licenses, or portions of digital collections which have simply not been sufficiently processed or analyzed.

The AFF format used for the M57 Patents corpus provides a number of features to support privacy-preserving views into raw data. Currently, there is both an application programming interface (API) and application-level (UNIX, Macintosh, and Windows) support for flexible access to disk images *without mounting the image on the host* or providing the raw, unencrypted bitstream to the end user. For example, an AFF file can be distributed in a container file with multiple streams of data that may be specified either as unencrypted or encrypted with unique keys and encryption schemes [7]. This allows a curator to specify *what* is provided for access, *how* it is provided, and *who* has permissions to access that material, all mediated by a single flexible container scheme.

4.2.2 Rights Management

As discussed above, it is often important to implement varying levels and types of access to data within repositories, based on rights associated with the materials. In the digital library literature, the most frequently discussed rights are those associated with intellectual property. However, there are a variety of other rights that one may need to address in the curation of digital collections, including cultural property, replevin and repatriation; right to privacy; protection of human subjects in research; privileged or protected information (e.g. client-attorney, healthcare, social services, library circulation, source-journalist); right to publicity; and prevention of misappropriation (including plagiarism).

For example, the M57 Patents corpus contains a significant amount of copyrighted materials, in the form of the Windows operating system. We therefore distribute two copies of the M57 Patents corpus: a version that contains the original disk images, and a second, derived corpus, in which all of the Microsoft executables have been “broken” so that they cannot be run by an end user. The original version is distributed as an encrypted volume and the key is only provided to organizations that have access to the Microsoft Developer Network (MSDN). The derived corpus is made freely available.

4.2.3 Dissemination Approaches

Facilitating access in the form of nuanced and tunable views into raw disk images requires tools and formats designed to provide a high degree of utility to the end user while maintaining the security of confidential, privileged, or commercially licensed data.

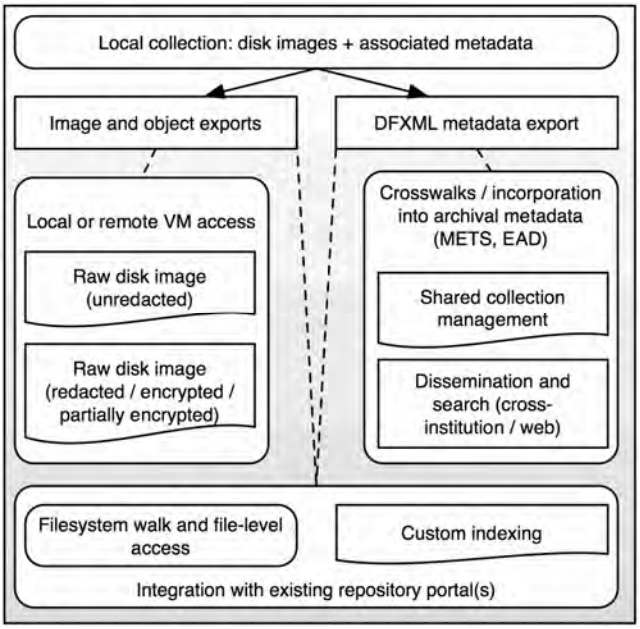

Figure 5: Data distribution; providing end-user access and supporting data reuse across organizations.

In cases when no privacy or intellectual property encumbrances are encountered, providing a complete disk image to users (either through virtualization or using software libraries to extract specific file system data) is a viable option. More frequently, it is necessary to provide only portions of the image. A variety of

approaches both for creating views into the raw data and extracting relevant selections of files and information to be stored for future use have been discussed in the forensics literature [15]. In other approaches, selective access to partial contents of a disk image are generated as needed using a pre-computed index, minimizing storage overhead while supporting high-performance access and retrieval [33][34].

A significant advantage of extracting data from a raw bitstream directly in order to create customized views is that the drive does not need to be mounted or searched as a live file system at the time of access. This reduces both processing overhead (as computationally-intensive virtualization processes are not necessary for many access events, such as viewing self-contained documents—although they can still be easily supported when necessary), ensures that the original bitstream is not changed, and simplifies secure handling of confidential and sensitive data. Redaction can be customized on a case-by-case basis depending on what a particular user is authorized to see.

In some cases, virtualization may be desirable or necessary. If the archivist has previously established that the raw disk image contains private or sensitive information, such access may be provided with reliable security in a number of ways. If the distribution mechanism in place requires serving out a complete object (rather than streamed network access), the AFF format can be used to create a clone of the original image for distribution in which specific byte sequences are encrypted. An alternate approach would be to create a "permissions overlay", through which file and folder level permissions are rewritten on demand (without altering the original bitstream) as portions of the image are streamed to a virtual machine client on the user's workstation.

4.3 Educational and Ethical Considerations

In addition to the logistical issues of applying digital forensics to the acquisition of materials, there are also deep and important institutional and ethical issues [24]. Many of these issues relate to adequate handling of data that should not be disclosed, or should be selectively disclosed/rendered. Because disk images are initially complete representations of an original physical device, this question is multifaceted.

First, curators of digital materials are responsible for understanding and implementing plans to handle "hidden" data. This is particularly relevant for disk images, because extensive investigation may be required to uncover contents of or fragments within a file system that would not appear in a simple walk of the directory tree.

Part of this responsibility hinges on commitments to donors and other stakeholders, who may specify data to be made available for access and depend on the curator to make informed judgments about collections which are not fully processed at the time of acquisition but may be stored in an unredacted form for some period of time.

Digital curation professionals must be able to use the tools and methods at their disposal to facilitate user inferences about the data with varying levels of certainty. In digital forensics, both professional practice and the software tools used to support investigations enable practitioners both to make such judgments and to provide some quantitative measures to explain the degree of support found in the data. There is great potential for collecting institutions to facilitate new forms of inferences, but this also carries potential responsibilities for ensuring that users of the materials do not assume an unrealistically high level of certainty based on available evidence [2].

There can be numerous sources of uncertainty related to the nature or sources of data. For example, parts of a page available through the WayBack Machine from the Internet Archive for a given date will not always accurately represent the parts of the page as available on that date [6]. Research on computing practices has also revealed many cases of shared computer use within a home [11], which introduces questions about the provenance of particular data on a computer. Carole Chaski has expressed many of these issues as "the keyboard problem," because there are limitations to how well one can determine "who was actually at the keyboard composing the document" [5]. When providing access to data from digital collections, it can be important for digital curation professionals to reflect to users of the collections that "all statements about digital states and events are hypotheses that must be tested to some degree" [4].

It is important that standards for admissibility and weighing of evidence in legal cases are stringent. In a legal context, these fundamental controls serve as a baseline to prevent miscarriages of justice. In other contexts there is a much broader sense of evidential value of archival materials [29]. Historians, genealogists, and other users of archival materials are quite accustomed to making inferences that can range from almost complete certainty to wild speculation, depending on the type and amount of available evidence.

Potentially valuable inferences can be posited at virtually every level of digital archival practice. Many practical advances come from relatively straightforward analyses of file formats, file contents, and file system metadata. Examples include names embedded in Microsoft Word documents (document author) to identify document templates and reuse, IP addresses associated with network activity conducted by a particular user on a particular system, and email addresses associated with particular user accounts and local client email databases or flat files.

Many components of this analysis can be automated. For example, cryptographic hash libraries can be constructed and used to determine the degree to which files may have been shared between systems or hard drives. MAC values stored by the file system can be used to determine document location changes and final edit dates and times.

5. CONCLUSION

We believe that mechanisms for capturing and providing access to disk images will play an increasingly important role in digital repositories. The basic technical strategies are already well-established both in digital forensics and for archives extracting data from fragile or aging media. In this work, we have addressed key aspects of handling disk images—imaging, selection and use of container formats, metadata extraction and production, provenance tracking, and rights management—that can help to support and extend the curation of digital collections.

By treating disk images as targets of digital curation activities, it is possible to expose avenues of information extraction, analysis, and distribution that are unavailable when the extensive information stored in the underlying file system is ignored or discarded. This does not replace or hinder document-centric archival activities. Nor does it eliminate the need for responsible

recordkeeping and digital curation practices both before and after the point of acquisition. Instead the curation of disk images can ensure retention and (when appropriate) use of *additional* information that is not visible in a simple file system view. Extracting, understanding, and analyzing this information can be complex. To facilitate the training and education of future digital archives experts, we have developed and deployed *realistic corpora* such as "M57 Patents," materials which simulate the information ecologies found in real-world digital computing environments.

We have presented work on the acquisition, analysis, and use of contextual information about digital objects necessary to support both archival and educational goals. By capturing information not only about the final representation of digital objects, but also about their representation and evolution on physical media and in their respective software environments, we show that previously unavailable methods of analysis and inference can be supported. We have explored some of the specific advantages of open source imaging, forensic acquisition, and forensic analysis tools, particularly the Advanced Forensic Format 4 and its associated libraries and utilities. We have described the creation of a novel realistic forensic dataset and its use as a training tool for digital archival education. We believe that the use of such corpora and associated forensics tools can fill a fundamental gap in current archival training and library and information science education.

6. ACKNOWLEDGMENTS

This work has been supported at the Naval Postgraduate School and UNC-CH by "Creating Realistic Forensic Corpora for Undergraduate Education and Research" (NSF Award DUE-0919593) and at UNC-CH by "Digital Acquisition Learning Laboratory" (Andrew W. Mellon Foundation).

7. REFERENCES

[1] Alink, W. 2005. *XIRAF: An XML-IR Approach to Digital Forensics*. Master's thesis, University of Twente, Enschede, The Netherlands, October 21.

[2] Caloyannides, M. A. 2006. Digital 'evidence' is often evidence of nothing. In *Digital Crime and Forensic Science in Cyberspace*, pages 334-339. Idea Group Pub., Hershey, PA.

[3] Caplan, Priscilla. 2009. *Understanding PREMIS: an overview of the PREMIS Data Dictionary for Preservation Metadata*. Library of Congress. February.

[4] Carrier, B. D. 2006. *A Hypothesis-Based Approach to Digital Forensic Investigations*. Doctoral Thesis. Purdue University.

[5] Chaski, C. 2007. The Keyboard Dilemma and Authorship Identification. In P. Craiger and S. Shenoi, editors, *Advances in Digital Forensics III: IFIP International Conference on Digital Forensics*, National Center for Forensic Science, Orlando, Florida, January 28-January 31. Volume 242 of IFIP International Federation for Information Processing. Springer, New York, NY.

[6] Cohen, F. 2008. Metrics for digital forensics. In Mini-MetriCon.

[7] Cohen, M.I., Garfinkel, S., and Schatz, B, 2009. Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information and Forensic Workflow. In *Proceedings of DFRWS 2009*. Montreal, Canada.

[8] Elford D., Pozo, N.D., Mihajlovic, S., Pearson, D., Clifton, G., and Webb, C. 2008. Media matters: developing processes for preserving digital objects on physical carriers at the National Library of Australia. In *74th IFLA General Conference and Council*, Quebec, Canada, August 10-14.

[9] Entlich, R. and Buckley, E. 2006. Digging up bits of the past: Hands-on with obsolescence. *RLG DigiNews*, 10(5).

[10] Farmer, D. and Venema, W. 2005. *Forensic Discovery*. Addison-Wesley, Upper Saddle River, NJ.

[11] Frohlich, D. and Kraut, R. 2003. The social context of home computing. In,*Inside the Smart Home*, pages 127-162. Springer, London.

[12] Garfinkel, S. and Cox, D. 2009. Finding and archiving the internet footprint. In *First Digital Lives Research Conference: Personal Digital Archives for the 21st Century*, London, UK, February 9-11.

[13] Garfinkel, S., Farrell, P., Roussev, V., and Dinolt, G. 2009. Bringing science to digital forensics with standardized forensic corpora. In *Proceedings of the 9th Annual Digital Forensic Research Workshop*, Montreal, Canada, August 17-19.

[14] Garfinkel, S. L. 2009. Providing cryptographic security and evidentiary chain-of-custody with the advanced forensic format, library, and tools. *International Journal of Digital Crime and Forensics*, 1(1):1-28, January-March.

[15] Garfinkel, S. L. and Shelat, A. 2003. Remembrance of data passed: A study of disk sanitization practices. *IEEE Security and Privacy*, 1, 1727.

[16] Garfinkel, S. 2009. Automating Disk Forensic Processing with SleuthKit, XML and Python, Systematic Approaches to Digital Forensics Engineering (IEEE/SADFE 2009), Oakland, California.

[17] Hedstrom, M.L., Lee, C.A., Olson, J.S. and Lampe, C.A. 2006. 'The old version flickers more': Digital Preservation from the User's Perspective. *American Archivist*, 69 (1). 159-187.

[18] John, J. L. 2008. Adapting existing technologies for digitally archiving personal lives: Digital forensics, ancestral computing, and evolutionary perspectives and tools. In *Proceedings of The Fifth International Conference on Preservation of Digital Objects (iPRES 2008)*, London, UK, September 29-30.

[19] John, J.L., Rowlands, I., Williams, P., and Dean, K. 2010. *Digital Lives: Personal Digital Archives for the 21st Century: an Initial Synthesis*. 2010.

[20] Kirschenbaum, M. G. 2008. *Mechanisms: new media and the forensic imagination*. MIT Press, Cambridge, MA.

[21] Kirschenbaum, M. G., Ovenden, R. and Redwine, G. 2010. *Digital Forensics and Born-Digital Content in Cultural Heritage Collections*. Council on Library and Information Resources, Washington, DC.

[22] Lavagnino, J. 1996. The analytical bibliography of electronic texts. In *Joint annual conference of the Association for*

Literary and Linguistic Computing and the Association for Computers and the Humanities, pages 180-182, Bergen, Norway.

[23] Lee, C. 2011. *A conceptual framework for contextual information in digital collections*. Journal of Documentation, 67, 1 (2011), 95-143.

[24] Lee, C. 2010., Automation in Digital Preservation - Computer-Supported Elicitation of Curatorial Intent., In *Dagstuhl Seminar Proceedings 10291*,

[25] McDonough, J.P. 2011. Packaging Videogames for Long-Term Preservation: Integrating FRBR and the OAIS Reference Model. *Journal of the American Society for Information Science and Technology*, 62 (1). 171-184.

[26] Oltmans, E., Van Diessen, R.J. and Wijngaarden, H.V. 2004. Preservation Functionality in a Digital Archive. In *Proceedings of the Fourth ACM/IEEE Joint Conference on Digital Libraries (JCDL 2004)*. Tucson, Arizona, June 7-11, 2004, ACM Press, New York, NY, 2004, 279-286.

[27] Rosenthal, D.S.H. 2010. *Format obsolescence: assessing the threat and the defenses*. Library Hi Tech, 28(2), 195-210.

[28] Ross, S. and Gow, A. 1999. *Digital archaeology: Rescuing neglected and damaged data resources*. Technical Report British Library Research and Innovation Report 108, British Library, London, February.

[29] Schellenberg, T.R. 1956. *The appraisal of modern records.* Bulletins of the National Archives, 8, October.

[30] Thibodeau, K. 2002. Overview of technological approaches to digital preservation and challenges in coming years. In *The State of Digital Preservation: An International Perspective*, pages 4-31. Council on Library and Information Resources.

[31] Underwood, W.E. and Laib, S.L. 2007. PERPOS: An Electronic Records Repository and Archival Processing System. *International Symposium on Digital Curation (DigCCurr 2007)*, Chapel Hill, NC, April 18-20.

[32] Underwood, W., Hayslett, M., Isbell, S., Laib, S., Sherrill, S., and Underwood, M. 2009. *Advanced Decision Support for Archival Processing of Presidential Electronic Records: Final Scientific and Technical Report*. Technical Report ITTL/CSITD 09-05. October.

[33] Woods, K. and Brown, G. 2008. Migration performance for legacy data access. *International Journal of Digital Curation*, 3(2), 74-88.

[34] Woods, K. and Brown, G. 2009. From imaging to access - effective preservation of legacy removable media. In *Archiving 2009: Preservation Strategies and Imaging Technologies for Cultural Heritage Institutions and Memory Organizations: Final Program and Proceedings*, pages 213-218. Society for Imaging Science and Technology, Springfield, VA.

[35] Woodyard, D. 2001. Data recovery and providing access to digital manuscripts. In *Information Online 2001 Conference*, Syndney, Australia, January 16-18.